Biophysical Journal Volume 86 February 2004 1105–1117

Interhelical Angle and Distance Preferences in Globular Proteins

Sangyoon Lee and Gregory S. Chirikjian
Department of Mechanical Engineering, Johns Hopkins University, Baltimore, Maryland

ABSTRACT Orientational preferences between interacting helices within globular proteins have been studied extensively over the years. A number of classical structural models such as "knobs into holes" and "ridges into grooves" were developed decades ago to explain perceived preferences in interhelical angle distributions. In contrast, relatively recent works have examined statistical biases in angular distributions which result from spherical geometric effects. Those works have concluded that the predictions of classical models are due in large part to these biases. In this article we perform an analysis on the largest set of helix-helix interactions within high-resolution structures of nonhomologous proteins studied to date. We examine the interhelical angle distribution as a function of spatial distance between helix pairs. We show that previous efforts to normalize angle distribution data did not include two important effects: 1), helices can interact with each other in three distinct ways which we refer to as "line-on-line," "endpoint-to-line," and "endpoint-to-endpoint," and each of these interactions has its own geometric effects which must be included in the proper normalization of data; and 2), all normalizations that depend on geometric parameters such as interhelical angle must occur before the data is binned to avoid artifacts of bin size from biasing the conclusions. Taking these two points into account, we find that there are very pronounced preferences for helices to interact at angles of approximately ± 160 and $\pm 20°$ in the line-on-line case. This pattern persists when the closest α-carbons in the helices vary from 4 to 12 Å. The endpoint-to-line and endpoint-to-endpoint cases also exhibit distinct preferences when the data is normalized properly. Analysis of the local structural interactions which give rise to these preferences has not been studied here and is left for future work.

INTRODUCTION

The protein folding problem has been a central topic in biophysics and structural biology for more than a quarter century (Anfinsen, 1973; Creighton, 1992). A number of ab initio methods for predicting the fold of a protein have been proposed (Srinivasan and Rose, 1995; Bonneau and Baker, 2001). And although it is believed that the principles driving protein folding are known (Baldwin and Rose, 1999a,b), the issue of exactly what chemical potentials to use to capture the behavior of proteins has been the subject of some debate. Proposed potentials have ranged from all-atom empirical models and explicit solvent (Weiner and Kollman, 1981) to those in which the mediating effects of solvent are built in implicitly (Maiorov and Crippen, 1992; Cheung et al., 2002), and to those derived from coarse-grained analysis of structures deposited in the Protein Data Bank (Miyazawa and Jernigan, 1985, 1996; Sippl, 1995). Any of these potentials can then be used together with energy minimization, conformational sampling, or dynamics techniques (Brooks et al., 1983; Skolnick and Kolinski, 1999; Abagyan, 1993) to try to predict the fold of a protein.

Although methods for protein secondary structure prediction are relatively reliable, developing methods for tertiary structure prediction based on first principles remains a challenging topic. Determining how elements of secondary structure assemble into proteins is therefore a critical intermediate step in solving the folding problem.

Given that the α-helix is a common and well-characterized secondary structure, and motifs built from α-helices form key elements of protein tertiary structure, many researchers have sought to determine principles for predicting the aggregation and contact patterns in α-helices (Crick, 1953; Levitt and Chothia, 1976; Richmond and Richards, 1978; Chothia et al., 1977, 1981; Finkelstein and Ptitsyn, 1987; Murzin and Finkelstein, 1988). Classical works have sought to explain helix-helix packing angle preferences in proteins based on models of steric fit and optimal packing of helices around hydrophobic cores. These models include ''knobs-into-holes'' (Crick, 1953), ''ridges-into-grooves'' (Chothia et al., 1977, 1981), and ''polyhedral helix globule'' (Finkelstein and Ptitsyn, 1987; Murzin and Finkelstein, 1988). In contrast, recent works have analyzed entries in the Protein Data Bank (PDB) (Berman et al., 2000) to look for patterns in the way helices interact in globular and membrane proteins (Lin et al., 1995; Lesk, 2001; Adamian and Liang, 2001; Bowie, 1997a,b; Fleming and Richards, 2000; Fleishman and Ben-Tal, 2002). Other works have compared the forces which stabilize globular and membrane proteins as a way to predict their assembly (Eilers et al., 2002; Walther et al., 1996; Robinson and Sligar, 1993; Efimov, 1979; Weaver, 1992; MacKenzie and Engelman, 1998; Zhou et al., 2000).

Database-driven approaches have the appeal that one can examine helix-helix pairs in a very large set of proteins, examine their crossing angle, and presumably make predictions based on these observations. Recently, however, several works have modeled the inherent statistical bias in

Submitted May 29, 2003, and accepted for publication October 1, 2003.

Address reprint requests to Gregory S. Chirikjian, Dept. of Mechanical Engineering, Johns Hopkins University, Baltimore, MD 21218. Tel.: 410-516-7127; Fax: 410-516-7254; E-mail: gregc@jhu.edu.

Sangyoon Lee's present address is School of Mechanical and Aerospace Engineering, Konkuk University, Seoul, South Korea.

0006-3495/04/02/1105/13 $2.00

distributions of interhelical angle due to spherical geometric effects (Bowie, 1997a) and fundamental differences in interactions between infinite and finite helix axes (Walther et al., 1998). Other works have examined the distribution of interhelix distance in interacting pairs (Reddy and Blundell, 1993). Most recently, a new analysis of helix-helix angle preferences has been performed in Trovato and Seno (2003). However, to our knowledge no prior work has investigated the joint distribution of interhelix distance and angle in proteins.

Bowie (1997a) and Walther et al. (1998) have shown in recent articles that several effects of orientational statistics naturally bias the number of observed structures to be greatest when α-helices cross near right angles. That is not to say that a peak in angular distributions of interacting helices is observed at $\pm 90^\circ$, but rather that observations without the proper normalization are biased toward those angles. In both Bowie (1997a) and Walther et al. (1998) it is reasoned that if one bins helix-helix angles to form a histogram, that this histogram should then be normalized by a histogram generated by all random noninteracting helix pairs to get an unbiased histogram. The core ideas in these articles are significant contributions to the statistical analysis of structural data in the PDB, though as shall be explained shortly, the particular implementations (and hence the resulting conclusions) must be reexamined for several reasons. In particular, while the amount of data available on helix-helix pairs in the PDB is substantial, it is not sufficient to generate robust histograms with very small bins. In fact, in Bowie (1997a) and Walther et al. (1998) the bin size used is 10° to generate the histogram of interhelical angle before normalization. Those works then normalize this histogram by the histogram generated by the bias (which is of the form $\sin \beta$ and $\sin^2 \beta$, in those articles, respectively). The problem with this approach is that it depends on the size of the bins used. As we shall show, this leads to the incorrect inference that antiparallel helix packings at 180° are preferred when the bias is removed.

Our modifications to the conceptual contributions in Bowie (1997a) and Walther et al. (1998) are that: 1), the correct normalization should be applied to each measurement before it is binned rather than after; and 2), all possible types of interactions (not only line-on-line) should be captured. Whereas no distinction between contact classes was made in Bowie (1997a), only line-on-line contacts were considered in Walther et al. (1998). There are two other contact classes which must be considered: endpoint-to-line and endpoint-to-endpoint. Each of these classes requires its own different normalization, and each class should be treated separately from the others. A detailed explanation of the importance of the effects of proper normalization and binning is given in the Appendix.

In addition, each measurement should be replaced with a probability density function (we use a Gaussian kernel) to account for measurement error to avoid binning artifacts altogether. A second modification results from the fact that there is another bias that should be removed which has not been observed in previous articles. Namely, the number of ways that helices can interact depends on the distance separating them, and therefore distance-dependent biases must be removed in addition to the previously observed orientation-dependent biases. All of these biases must be negated before binning is performed. In principle, if the amount of available experimental data were tremendously larger than it is, it would be possible to make very fine bins and follow the procedures outlined in Bowie (1997a) and Walther et al. (1998), but with the amount of data currently available, normalization must precede binning to avoid artifacts due to the size and location of bins.

The emphasis in our article is different for several reasons, including those listed above. In part this is because we are interested in examining how interhelical angle distributions vary with the relative spatial distance between the helices. In other words, one of our goals is to determine how persistent the angle preferences are as the spatial distance between helices is varied. We also explain statistical effects not accounted for in prior works that reveal clear preferences in helix-helix angles.

METHODS

We examine 1290 protein structures which have been resolved to 2.0 Å or better and possess $<20\%$ of their sequences in common. This data is obtained from Wang and Dunbrack (2003). Within these structures there are 12,207 helices (the vast majority of which are α-helices) and there are 90,438 helix-helix pairs (many of which do not represent helices in direct contact). We examine the distribution of helix-helix angle over all of these pairs. The full angular distribution is broken down according to the spatial distance that separates helices from each other. Interhelical distance is measured in two ways: 1), along the shortest line segment connecting each finite helix axis; and 2), by finding the minimal distance between every pair of α-carbons in the two helices under consideration. These two methods for measuring interhelical distance are illustrated in Fig. 1.

The angle distributions examined here are over the range of $(-180, 180^\circ)$, corresponding to helix axes with directionality. We number residues and helices sequentially in the usual way, starting at the N-terminus. We assign a unit vector along the helix axis (choosing the direction representing an increase in sequence number over the direction representing decrease in sequence number). The angle between two helices is considered to be positive if helix 2 is rotated clockwise relative to helix 1 by an angle between 0 and 180° about the unique line segment pointing from the axis of helix 1 to that of helix 2 and intersects both axes at right angles. This convention does not depend on the numbering of the helices after the direction of the helix unit

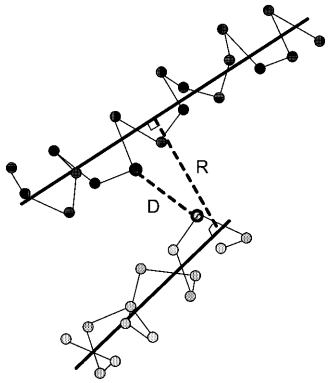

FIGURE 1 Definition of interhelical distance, R, between helix axes and distance, D, between helices as measured between closest α-carbons.

vector is established. Such a line of closest approach will always exist for infinitely long helix axes. In contrast, α-helices have finite length, and so a distinction must be made between the axis of the helix and that finite part of the axis which lies inside the helix. We reiterate the distinction made in Walther et al. (1998) and call the ideal case an *infinite axis* and the actual case a *finite axis*. There are three very different scenarios that are possible: 1), the line of closest contact between infinite axes intersects both finite axes at right angles, in which case a line-on-line contact is made; 2), the line of closest contact between two finite axes meets one helix at its end and the other at a right angle, in which case an endpoint-to-line contact results; and 3), both helices interact only at their ends, in which case an end-to-end contact results.

We break the discussion into the three cases. In the case when the helices interact by crossing such that the line of closest approach in Fig. 1 intersects the interior of both finite axes, then the distances R and D are on average related as $R = D + 2a$ where a is the average radius of an α-helix (as measured from its axis to the α-carbons). As has been observed (Reddy and Blundell, 1993), α-helices interact over a range of interhelical distances.

An implicit assumption in the study of helix-helix interactions is that α-helices are essentially rigid objects. The relative position and orientation between rigid bodies can be expressed with a pair $(A, \mathbf{a})$ where A is a rotation matrix and $\mathbf{a}$ is a translation vector. To be more precise, assume two bodies have reference frames attached to them in some canonical way. We define the relative motion that will take the frame of reference attached to body 1 into the frame of reference attached to body 2 to be $(A, \mathbf{a})$. Then from the perspective of body 2, the relative motion that it would need to undergo for its frame to become coincident with frame 1 would be $(A^{\mathrm{T}}, -A^{\mathrm{T}}\mathbf{a})$. The set of all such rigid-body motions forms a manifold, and the operation of composing two rigid-body motions endows this manifold with the structure of a Lie group. It is convenient to think of this Lie group as the set of all 4×4 *homogeneous* transformation matrices of the form

$$H = \begin{pmatrix} A & \mathbf{a} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{pmatrix}, \tag{1}$$

in which case, matrix multiplication corresponds to the group operation and

$$\begin{aligned} H[A_1, \mathbf{a}_1]H[A_2, \mathbf{a}_2] &= H[(A_1, \mathbf{a}_1)\circ(A_2, \mathbf{a}_2)] \\ &= H[(A_1A_2, A_1\mathbf{a}_2 + \mathbf{a}_1)]. \end{aligned}$$

Two special kinds of homogeneous transformations are pure rotations and pure translations along the axes of local coordinate systems,

$$rot(\mathbf{e}_i, \theta) = \begin{pmatrix} R_i(\theta) & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{pmatrix}$$

and

$$trans(\mathbf{e}_i, x) = \begin{pmatrix} I & x\mathbf{e}_i \\ \mathbf{0}^{\mathrm{T}} & 1 \end{pmatrix}.$$

It is well-known (see Chirikjian and Kyatkin, 2001) that the Lie group of rigid-body motions in three-dimensional space possesses a unique bi-invariant integration measure. That is, there is only one correct way to integrate over rigid-body motions. In particular, given a function $f(A, \mathbf{a})$ describing the relative pose (position and orientation) of the frame of reference attached to body 2 relative the frame of reference attached to body 1, if the parameters defining A and those defining $\mathbf{a}$ are independent, then there is only one correct way to integrate it as

$$I = \int_{A\in SO(3)} \int_{\mathbf{a}\in R^3} f(A, \mathbf{a}) d\mathbf{a}\, dA.$$

Here $SO(3)$ is the group of rotations in three-dimensional space and dA is its bi-invariant integration measure. If $A = A(\alpha, \beta, \gamma)$ is the common ZXZ Euler-angle parameterization, then (Chirikjian and Kyatkin, 2001):

$$dA = \frac{1}{8\pi^2} \sin\beta\, d\alpha\, d\beta\, d\gamma.$$

It is this orientational volume element (which is fundamentally the same as that for the unit sphere) which, by itself, leads to orientational normalizations such as in Bowie (1997a). In contrast, if the position of the origin of frame 2 is described in Cartesian coordinates relative to frame 1, then

$$d\mathbf{a} = dx\, dy\, dz.$$

For the three cases shown in Fig. 2, spatial rigid-body motions of helix 2 relative to helix 1 are parameterized in three different ways. And it is not obvious a priori what the correct integration measure should be as a function of the parameters describing each of those models. Determining this is essential to correctly account for the statistical biases inherent in the three data sets. For this reason, the general method for determining the volume element for integrating over rigid-body motions is derived here. The results for all

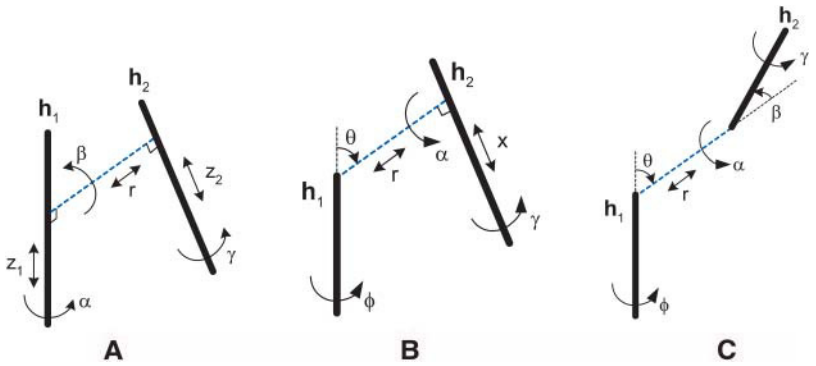

FIGURE 2 Helix-helix interaction diagrams. (*a*) Case 1: line-to-line. (*b*) Case 2: endpoint-to-line. (*c*) Case 3: endpoint-to-endpoint.

three cases in Fig. 2 are then given. The explicit calculations are contained in the Appendix.

For ''small'' rigid-body motions,

$$H \approx I + \begin{pmatrix} \Omega & \mathbf{v} \\ \mathbf{0}^{\mathrm{T}} & 0 \end{pmatrix} \Delta t, \tag{2}$$

where the matrix Ω is skew-symmetrically defined as $\Omega = -\Omega^{\mathrm{T}}$. It describes an infinitesimal orientational displacement. In fact, the angular velocity vector ω can be extracted from the matrix Ω to describe the rotational part of the displacement as

$$\mathrm{vect}(\Omega) = \boldsymbol{\omega}.$$

That is, if

$$\Omega = \begin{pmatrix} 0 & -w_3 & w_2 \\ w_3 & 0 & -w_1 \\ -w_2 & w_1 & 0 \end{pmatrix}, \quad \boldsymbol{\omega} = \begin{pmatrix} w_1 & w_2 & w_3 \end{pmatrix}^{\mathrm{T}}.$$

Since the second term in Eq. 2 consists mostly of zeros, it is common to extract the information necessary to describe the motion as

$$\begin{pmatrix} \Omega & \mathbf{v} \\ \mathbf{0}^{\mathrm{T}} & 0 \end{pmatrix}^{\vee} = \begin{pmatrix} \boldsymbol{\omega} \\ \mathbf{v} \end{pmatrix}. \tag{3}$$

This six-dimensional vector is called an *infinitesimal screw motion* or *infinitesimal twist*.

Given a homogeneous transform consisting of motions that are not necessarily small,

$$H(\mathbf{q}) = \begin{pmatrix} A(\mathbf{q}) & \mathbf{a}(\mathbf{q}) \\ \mathbf{0}^{\mathrm{T}} & 0 \end{pmatrix},$$

parameterized with coordinates $(q_1, \ldots, q_6)$, which we write as a six-dimensional vector $\mathbf{q}$, one can express the homogeneous transform corresponding to a slightly changed set of parameters as the truncated Taylor series

$$H(\mathbf{q} + \delta\mathbf{q}) = H(\mathbf{q}) + \sum_{i=1}^{6} \delta q_i \frac{\partial H}{\partial q_i}(\mathbf{q}). \tag{4}$$

This result can be shifted to the identity transformation by multiplying on the left by H^{-1} to define an equivalent relative infinitesimal motion. In this case we write

$$\begin{pmatrix} \boldsymbol{\omega}_{\mathrm{R}} \\ \mathbf{v}_{\mathrm{R}} \end{pmatrix} = \vartheta_{\mathrm{R}}(\mathbf{q})\dot{\mathbf{q}} \quad \text{where}$$

$$\vartheta_{\mathrm{R}}(\mathbf{q}) = \left[\left(H^{-1} \frac{\partial H}{\partial q_1} \right)^{\vee}, \cdots, \left(H^{-1} \frac{\partial H}{\partial q_6} \right)^{\vee} \right]. \tag{5}$$

Here $\vartheta_{\mathrm{R}}(\mathbf{q})$ is a 6×6 matrix, and the spatial velocity $\mathbf{v}_{\mathrm{R}}$ and special angular velocity $\boldsymbol{\omega}_{\mathrm{R}}$ are defined as

$$\mathbf{v}_{\mathrm{R}} = A^{\mathrm{T}}\dot{\mathbf{a}} \quad \text{and} \quad \boldsymbol{\omega}_{\mathrm{R}} = \mathrm{vect}(A^{\mathrm{T}}\dot{A}).$$

The unique volume element for correctly integrating over rigid-body motions in the coordinates $q_1, \ldots, q_6$ is (Chirikjian and Kyatkin, 2001),

$$dH = |\det \vartheta_{\mathrm{R}}| dq_1 \ldots dq_6. \tag{6}$$

If three parameters are used to describe orientation, and three are used to describe position, then Eq. 6 reduces to the product of positional and orientational volume elements discussed earlier. However, in the cases shown in Fig. 2, the six parameters describing the pose of one line segment relative to another cannot be decoupled into those which independently describe position and orientation. Hence, Eqs. 5 and 6 must be computed explicitly to determine the proper normalization of data in each case, as shown in Fig. 2.

In cases where the two rigid bodies have symmetries (as is the case for a line segment), and hence the function $f(A, \mathbf{a})$ is constant over certain coordinates, it makes sense to use a parameterization which captures this fact, and then integrate out all such coordinates. In this way marginal probability densities on a space of reduced dimension can be examined. Below, the form of the volume elements is given, and the proper reductions are performed for the three cases shown in Fig. 2.

Case 1: Line-to-line interaction

In Fig. 2 *a*, the series of rigid-body motions that result in the frame attached at the base of helix 1 being moved to the base of helix 2 parameterize the homogeneous transformation of

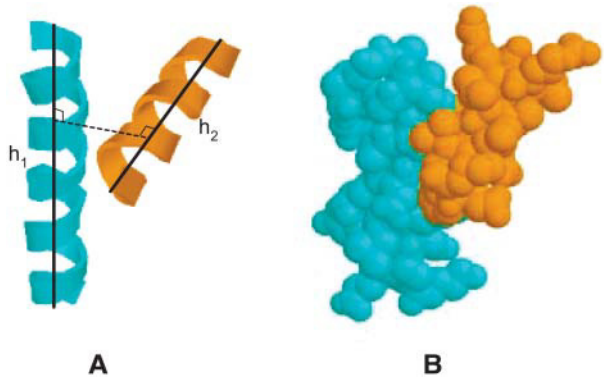

FIGURE 3 Case 1: Helix-helix interactions. (*a*) Ribbon representation. (*b*) All-atom. The helices in this figure are the third helix (21 residues with sequence numbers 60–80) and the fifth helix (14 residues with sequence numbers 93–106) in a protein with ID 119L.

$$H(\alpha, z_1, \beta, r, \gamma, z_2) = rot(\mathbf{e}_3, \alpha)trans(\mathbf{e}_3, z_1)rot(\mathbf{e}_1, \beta) \times trans(\mathbf{e}_1, r)rot(\mathbf{e}_3, \gamma)trans(\mathbf{e}_3, z_2). \quad (7)$$

Substitution into Eq. 5, and following the calculations in Appendix A, this results, to within an arbitrary multiplicative constant, in

$$|\det \vartheta_R| = \sin^2 \beta. \quad (8)$$

This is the same as the normalization obtained in Walther et al. (1998). Fig. 3 shows an example of such a pair.

Case 2: End-to-line interaction

Observing Fig. 2 *b*, it is clear that the rigid-body motion taking frame 1 into frame 2 is of the form

$$H(\phi, \theta, r, \alpha, x, \gamma) = rot(\mathbf{e}_3, \phi)rot(\mathbf{e}_1, \theta)trans(\mathbf{e}_3, r)rot(\mathbf{e}_3, \alpha) \times trans(\mathbf{e}_1, x)rot(\mathbf{e}_1, \gamma). \quad (9)$$

Following the calculations in Appendix A,

$$|\det \vartheta_R| = r \sin \theta. \quad (10)$$

Fig. 4 shows an example of such a pair.

Case 3: End-to-end interaction

Looking at Fig. 2 *c*, the sequence of concatenated rigid-body motions that takes frame 1 to frame 2 is

$$H(\phi, \theta, r, \alpha, \beta, \gamma) = rot(\mathbf{e}_3, \phi)rot(\mathbf{e}_1, \theta)trans(\mathbf{e}_3, r) \times rot(\mathbf{e}_3, \alpha)rot(\mathbf{e}_1, \beta)rot(\mathbf{e}_3, \gamma). \quad (11)$$

Following the calculations in Appendix A,

$$|\det \vartheta_R| = r^2 \sin \beta \sin \theta. \quad (12)$$

Fig. 5 shows an example of such a pair.

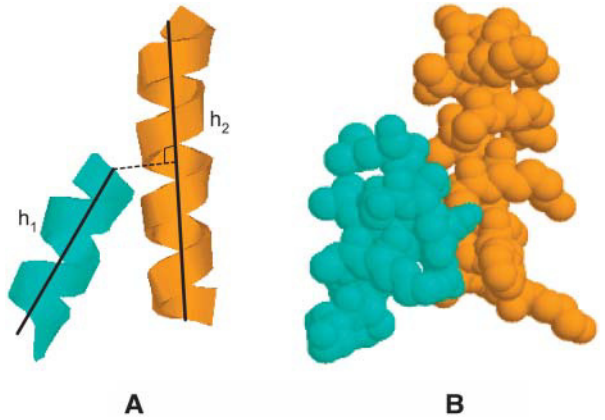

FIGURE 4 Case 2: Helix-helix interactions. (*a*) Ribbon representation. (*b*) All-atom. The helices in this figure are the fifth helix (11 residues with sequence numbers 93–103) and the 18th helix (17 residues with sequence numbers 351–367) in a protein with ID 16PK.

We consider two helices to be a candidate interacting pair if the distance D shown in Fig. 1 is within 15 Å. This criterion is somewhat different than previous works in which interacting helices are defined as those for which at least several atoms from one helix are in contact with those of the other. The reason for our choice is that having atoms in contact is neither a necessary nor a sufficient condition for determining the orientation between helices. In part this is because of the articulate nature of side chains and in part because of the long-range effects of certain kinds of molecular forces. Our cutoff of 15 Å was imposed after examining all 90,438 helix pairs using the statistical methods described below, and determining that for all helices outside of this distance cutoff there is no orientational order between helices, whereas below this threshold there is orientational order. Since in the next section we display helix interaction data as a function of both interhelical angle and distance, the criteria used for defining interacting pairs is somewhat more

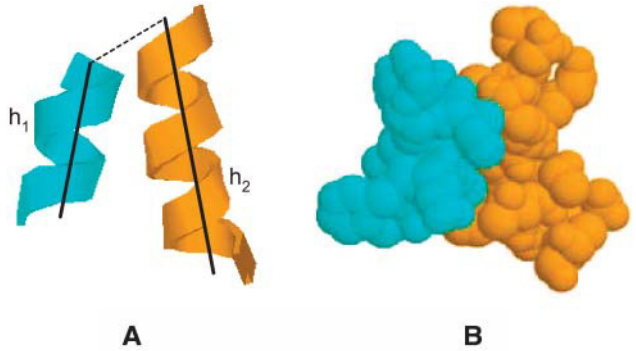

FIGURE 5 Case 3: Helix-helix interactions. (*a*) Ribbon representation. (*b*) All-atom. The helices in this figure are the first helix (nine residues with sequence numbers 3–11) and the 10th helix (12 residues with sequence numbers 143–155) in a protein with ID 119L.

flexible than when one-dimensional angular histograms are used; if our cutoff criteria were too loose, then the resulting two-dimensional plots would have large areas with no peaks, and if the cutoff criteria were too severe then it would be clear by looking at the plots that peaks would be clipped.

The effects of measurement error

Two potential sources of error can be introduced in our analysis of helix-helix interactions: 1), it is possible for a helix pair to be misclassified; and 2), the measurement of helix-helix angles is sensitive to the method used to define the helix axes. Here we describe statistical techniques that reduce the sensitivity of computed distributions to these phenomena. A very different approach to handling measurement errors is described in Trovato and Seno (2003).

The three distinct classes for helix-helix interaction described earlier form a partition of the six-dimensional space of rigid-body motions into three disjoint regions. Within each of the resulting six-dimensional regions the given parameterizations hold. If it were possible to exactly define and measure the endpoints of the finite axis of each α-helix in the PDB and if the backbone of every α-helix observed ideal geometry, then the observed six-dimensional data describing the relative pose of every pair of helices could be normalized directly using the given factors. However, in reality measurement errors will always exist. For helix pairs that interact in a manner which is on the border between any two of the different interaction classes, it is possible that such pairs can be binned incorrectly. For this reason, the most rigorous treatment would treat each observation as a Gaussian distribution on the six-dimensional space of rigid body motions. In this way, the effects of an observation in one interaction class can be allowed to bleed into others. In borderline cases, this would reduce errors introduced by an all-or-nothing classification of each observed helix-helix pair. Although the concept of a Gaussian (or heat) kernel for the group of rigid-body motions exists (Chirikjian and Kyatkin, 2001), it is somewhat involved to implement. We have therefore taken the time-consuming approach of examining borderline cases and convincing ourselves that they have been classified correctly.

The second source of error (due to sensitivity in the definition of the helix axis) means that even when there is confidence in the class of interaction, the exactness of observed parameters such as angles and distances may be in question. As an example, suppose one is interested in the one-dimensional distribution of angle in the line-on-line case. Then, given a set of interhelical angles in the line-on-line case $\beta_1, \ldots, \beta_N\}$, one wishes to construct an estimate of the underlying probability density that describes the distribution from which these values are drawn. If every measurement were exact, and $N \to \infty$, then in principle this distribution could be constructed as

$$f(\beta) = \frac{1}{N}\sum_{i=1}^{N}\frac{\delta(\beta-\beta_i)}{\sin^2\beta} = \frac{1}{N}\sum_{i=1}^{N}\frac{\delta(\beta-\beta_i)}{\sin^2\beta_i} = \frac{1}{N}\sum_{i=1}^{N}\frac{\delta(\beta-\beta_i)}{\int_{-\pi}^{\pi}\delta(\beta'-\beta_i)\sin^2\beta'\,d\beta'},$$

where $\delta(\beta)$ is the Dirac delta function.

However, proteins are dynamic objects and each crystal structure provided in the PDB only represents a best estimate of an average over an ensemble of many similar (but not exactly the same) structures. In addition to thermal fluctuations, factors such as the specific refinement program that is used and the resolution of the structure all come into play in adding some uncertainty in the structures reported in the PDB. To make matters worse, there is no unique way to define the axis of an α-helix within a protein. One could fit the best-fit line to all or some windowed segment of α-carbons. One could use all the atoms in the helix. One could construct either local helix axes or averaged local helix axis. Or one could construct the best-fit least-squares curve to describe a bent/curved/supercoiled helix and define the interhelix axis in terms of tangents to these curves.

Due to uncertainties in PDB structures and inconsistencies in the way helix angles are measured, as well as the fact that N is finite, one would like to replace the above equation with

$$\tilde{f}(\beta,t) = \frac{1}{N}\sum_{i=1}^{N}\frac{k(\beta-\beta_i,t)}{\int_{-\pi}^{\pi}k(\beta'-\beta_i,t)\sin^2\beta'\,d\beta'},$$

where $k(\beta,t)$ is a kernel chosen by the user to construct $\tilde{f}(\beta)$ such that $||\tilde{f}-f|| \to 0$ as the accuracy and number of measurements both increase. A natural kernel in the current context is that obtained by ''wrapping'' the solution of the heat equation on the line,

$$h(x,t) = \frac{1}{\sqrt{2\pi t}}e^{-x^2/2t},$$

around the circle to obtain

$$k(\beta,t) = \sum_{n=-\infty}^{\infty} h(\beta-2\pi n,t).$$

When t is very small (corresponding to high confidence in the measurement) the above is accurately approximated when only the $n = 0$ term is retained.

In general, the quantity

$$C(\beta_i,t) = \int_{-\pi}^{\pi}k(\beta'-\beta_i,t)\sin^2\beta'\,d\beta'$$

can be computed in closed form as

$$C(\beta_i,t) = \frac{1}{2}(1-e^{-2t}\cos 2\beta_i),$$

and for $t > 0$ this provides some regularization which reduces the sensitivity of dividing directly by $\sin^2\beta_i$ when there are uncertainties in the value β_i. This is particularly important when β_i is near the singularities (0, $\pm$180°).

Normalization of marginal probability densities

Previously in this section, the weighting functions which would be used to normalize six-dimensional data on the frequency of occurrence of relative position and orientation of helices in the three different contact cases were examined. Since it is difficult to visualize six-dimensional data, and since the number of interacting helix pairs in high-resolution nonhomologous protein structures is too small to form robust statistics in such a high-dimensional space, it is advantageous to view spatial relationships described by marginal densities of the full pose distribution. Let Φ_1 and Φ_2, respectively, denote two distinct sets of pose parameters such that the sum of their dimensions is six. Let Φ_1 be the parameters which are the argument of the marginal distribution of interest, and Φ_2 be the set of variables which are integrated out of the pose distribution. If the unnormalized density describing the relative frequency of occurrence of observed pose is $f(\Phi_1, \Phi_2)$, and the normalized density (which is the one of interest) is $\rho(\Phi_1, \Phi_2)$, then writing $|\det \vartheta_R| = w(\Phi_1)w(\Phi_2)$, one observes

$$f(\Phi_1, \Phi_2) = w_1(\Phi_1)w_2(\Phi_2)\rho(\Phi_1, \Phi_2).$$

The observed marginal density (i.e., that which is formed by binning data in a grid of Φ_1 values without regard to the Φ_2 values) is then

$$f_1(\Phi_1) = \int f(\Phi_1, \Phi_2)d\Phi_2 = w_1(\Phi_1) \int \rho(\Phi_1, \Phi_2)w_2(\Phi_2)d\Phi_2.$$

By defining

$$\rho_1(\Phi_1) = \int \rho(\Phi_1, \Phi_2)w_2(\Phi_2)d\Phi_2,$$

it follows that the unbiased marginal density describing the preferred values of Φ_1 is obtained by normalizing the marginal observed density as

$$\rho_1(\Phi_1) = f_1(\Phi_1)/w_1(\Phi_1).$$

In other words, when normalizing observed marginal densities, it is not the full Jacobian determinant which is used as a normalization factor, but rather only that part of it which depends on the parameters which are the arguments of the marginal density.

For example, if in case 3 we want a two-dimensional contour plot of unbiased density on the domain parameterized by (r,β), then the normalization factor would be $r^2 \sin\beta$. In contrast, if we wanted a contour plot on (β,θ), the normalization would be $\sin\beta \sin\theta$, and if one wanted a contour plot in (α,γ), no normalization factor would be required.

In addition, the normalization for each case should be modified to account for errors as discussed in the previous subsection.

RESULTS

In what follows, data is normalized with the weighting functions computed in the previous section. In addition, each resulting marginal density is normalized to be a probability density so that the contour values are not sensitive to the number of recorded data points.

Case 1: Line-to-line interaction

The number of helix pairs in this case is 5534 pairs. Fig. 6 shows a normalized contour plot of relative frequency of occurrence of helix-helix interaxis angle β and distance D between nearest α-carbons.

A regularization parameter of $t = 0.01$ was chosen. A high peak at $\beta = 161.0°$ and $D = 5.1$ Å and three low peaks (at approximately $\pm 20°$ and $-165°$) are found. We also investigated the effects of increasing the value of t. For example, at the large value of $t = 1.0$, the peak at $-20°$ appears to melt away, and the peaks at $-165°$ and $+20°$, respectively, shift to $-125°$ and $+45°$, and both spread over a range of $\sim 30°$.

Note that the angle β is actually the same as the angle Ω in previous works (Bowie, 1997a; Walther et al., 1998). In agreement with Bowie (1997a) and Eilers et al. (2002) we find that antiparallel helix-helix interactions are more common and are more tightly packed than parallel interactions. Although much of the contribution to the major peak in Fig. 6 comes from helix pairs that are separated by a sequential distance of <10 residues, the locations of all the peaks appear to persist when a sequential cutoff of at least 20 intervening residues is imposed. In this case the relative height of the peaks is redistributed more evenly among them.

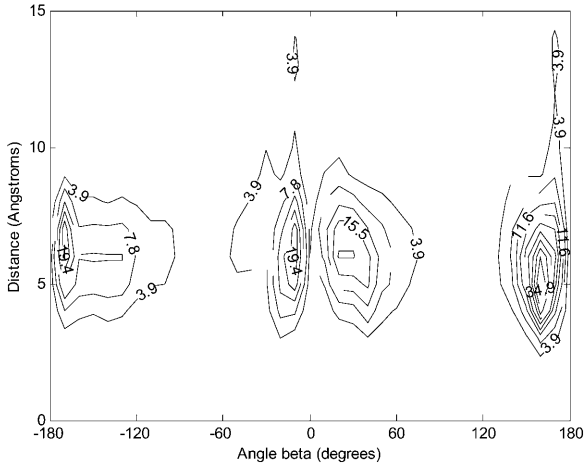

FIGURE 6 Case 1: Normalized contour plot of relative frequency of occurrence of helix-helix interaxis angle β and distance D between nearest α-carbons.

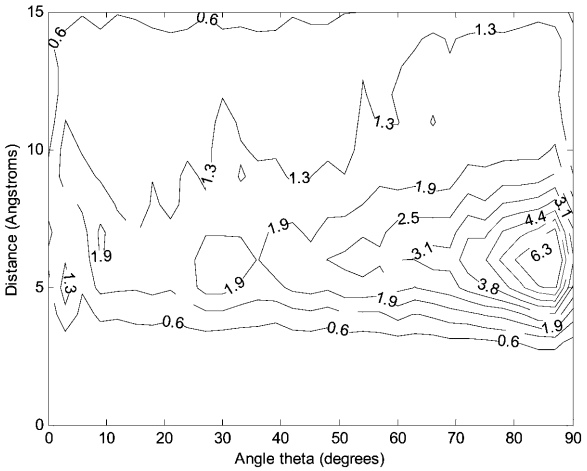

FIGURE 7 Case 2: Normalized contour plot of relative frequency of occurrence of angle θ and distance D between nearest α-carbons.

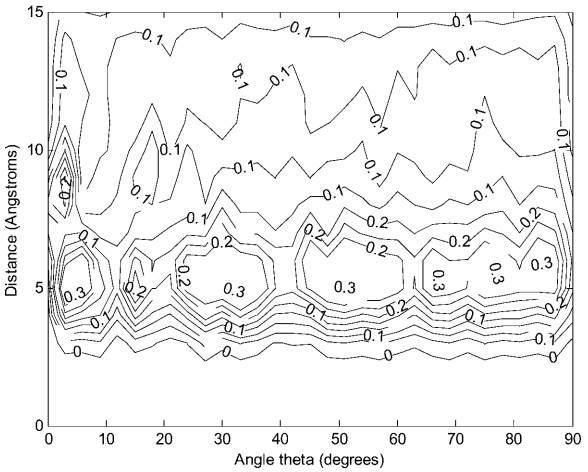

FIGURE 9 Case 3: Normalized contour plot of relative frequency of occurrence of angle θ and distance D between nearest α-carbons.

Case 2: End-to-line interaction

This case has 13,984 helix-helix pairs. In Fig. 7 the relative frequency at which a high peak is found is at the angle $\theta = 85.6^\circ$ and the distance $D = 6.1$ Å. The location of this peak appears to persist regardless of the sequential distance between the helices (as measured by the number of intervening residues).

In case 2, other relationships can also be explored. For example, the relative frequency of occurrence of values for the angle α as a function of distance between closest α-carbons in interacting helices is examined in Fig. 8. In this figure several low peaks are found. Those which occur close to $\pm 180^\circ$ have major contributions from helices which are separated in sequence from each other by <10 intervening residues, whereas the other observed peaks occur for larger values of sequential distance. We also constructed and examined two-dimensional plots of α and θ without regard to interhelical distance, but found no significant peaks in this case.

Case 3: End-to-end interaction

The number of helix pairs in this case is 8847 pairs. Fig. 9 shows a normalized contour plot of relative frequency of occurrence of helix-helix interaxis angle θ and distance D between nearest α-carbons. The relationship between angle β and distance D is illustrated in Fig. 10. In both plots, no significant peak is found. However, in Fig. 11, a normalized contour plot, showing relationships between angle θ and angle β, one high peak at $\theta = 3.1^\circ$, $\beta = 6.8^\circ$, and several low peaks are found.

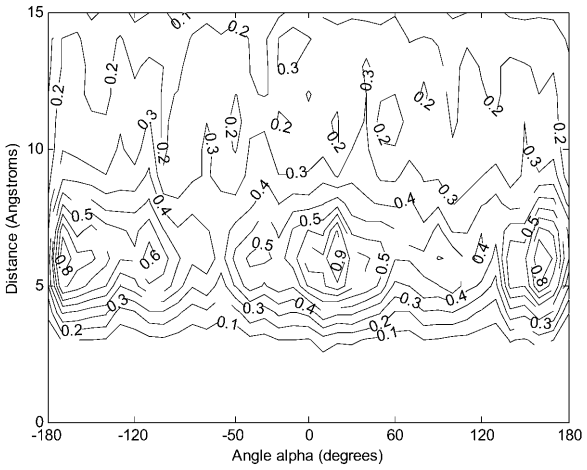

FIGURE 8 Case 2: Normalized contour plot of relative frequency of occurrence of angle α and distance D between nearest α-carbons.

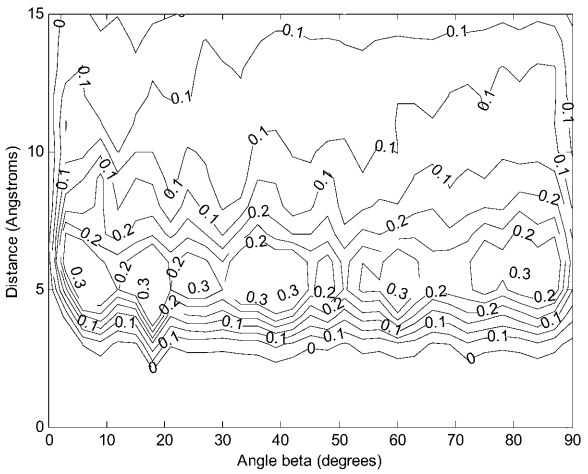

FIGURE 10 Case 3: Normalized contour plot of relative frequency of occurrence of angle β and distance D between nearest α-carbons.

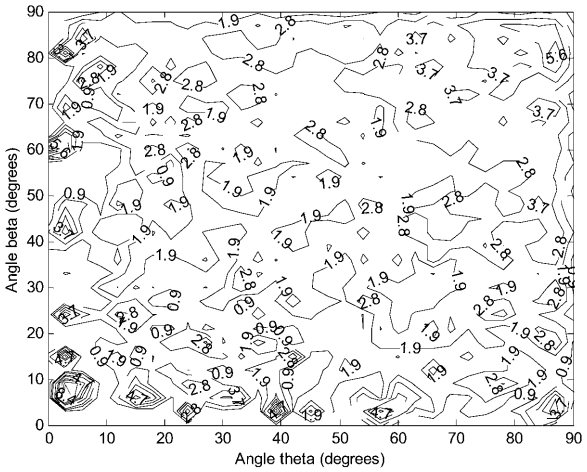

FIGURE 11 Case 3: Normalized contour plot of relative frequency of occurrence of angle θ and angle β.

CONCLUSIONS

Interhelical angle distributions have been studied as a function of spatial and sequential distance between helices in globular proteins. Spherical geometric effects due to both the distance and angle between interacting helices have been used to normalize the data. When normalized in this way, distinct preferences close to parallel ($\pm 20°$) and near antiparallel ($\pm 160°$) packings emerge in the line-to-line case. These four sets of values are consistent with the fact that if a helix is flipped end-over-end by 180°, the ridges/grooves are essentially the same as the original helix. It seems plausible that any subtle differences in ridges/grooves between a helix standing C to N vs. N to C may be accounted for by articulation of side chains. Therefore, 20–180 = −160 and 160–180 = −20 should be expected if 20 and 160 are expected. Such angles can be expected using reasoning similar to that behind the ridges-into-grooves formulation. In future work we plan to examine the local structural details which give rise to these preferences. We believe that empirically obtained helix-helix potentials obtained from these distributions may be useful for incorporation in protein folding algorithms.

APPENDIX: THE CORRECT NORMALIZATION FOR HELIX PAIR DATA

This Appendix consists of two parts. In the first part we provide the detailed calculations required to derive the correct normalization factors for helix-helix pair data. In the second, we show why binning before normalization can produce substantial artifacts which depend on bin size.

A.1 Derivation of correct normalization factors

In this Appendix section the computations resulting in the volume elements in the main part of the text are given in detail. Recall that for counter-clockwise rotations about the $\mathbf{e}_3$, $\mathbf{e}_2$, and $\mathbf{e}_1$ axes:

$$R_3(\phi)=\begin{pmatrix}\cos\phi & -\sin\phi & 0\\ \sin\phi & \cos\phi & 0\\ 0 & 0 & 1\end{pmatrix};\quad R_2(\phi)=\begin{pmatrix}\cos\phi & 0 & \sin\phi\\ 0 & 1 & 0\\ -\sin\phi & 0 & \cos\phi\end{pmatrix};$$

$$R_1(\phi)=\begin{pmatrix}1 & 0 & 0\\ 0 & \cos\phi & -\sin\phi\\ 0 & \sin\phi & \cos\phi\end{pmatrix}.$$

Each of these basic rotations can be written as the matrix exponential

$$R_i(\phi) = \exp(\phi E_i),$$

where

$$E_1 = \begin{pmatrix}0 & 0 & 0\\ 0 & 0 & -1\\ 0 & 1 & 0\end{pmatrix};\quad E_2 = \begin{pmatrix}0 & 0 & 1\\ 0 & 0 & 0\\ -1 & 0 & 0\end{pmatrix};$$

$$E_1 = \begin{pmatrix}0 & -1 & 0\\ 1 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}.$$

Multiplication of these matrices with any vector can be expressed as

$$E_i\mathbf{x} = \mathbf{e}_i \times \mathbf{x},$$

where $\mathbf{e}_i$ is the i^{th} natural unit basis vector in three-dimensional space. The above relationship is described using the notation $\mathbf{e}_i = \text{vect}(E_i)$, and since the ''vect'' operation is linear, it can be used to relate any three-dimensional vector and any 3 × 3 skew-symmetric matrix.

Note that $E_i\,\mathbf{e_i} = 0$ and $E_1\mathbf{e}_2 = \mathbf{e}_3$, $E_3\mathbf{e}_1 = \mathbf{e}_2$, and $E_2\mathbf{e}_3 = \mathbf{e}_1$. In contrast, $R_i(\phi)\ \mathbf{e_i} = \mathbf{e_i}$. Another property which is used multiple times to obtain the results presented below is that for any 3 × 3 rotation matrix R and any 3 × 3 skew symmetric matrix Ω,

$$\text{vect}(R\Omega R^{\mathrm{T}}) = R\,\text{vect}(\Omega).$$

Case 1: Line-to-line interaction

Performing the multiplications in Eq. 7, one finds

$$H=\begin{pmatrix}R_3(\alpha)R_1(\beta)R_3(\gamma) & z_2R_3(\alpha)R_1(\beta)\mathbf{e}_3 + \\ & rR_3(\alpha)\mathbf{e}_1 + z_1\mathbf{e}_3\\ \mathbf{0}^{\mathrm{T}} & 1\end{pmatrix}. \quad \text{(A1)}$$

Unlike the other two cases below, in this case three variables which appear in the translation part of the homogeneous transformation matrix do not appear in the rotation part. This gives a block structure to the Jacobian matrix, and makes the determinant easy to compute.

In particular, if we group the variables as $\mathbf{q}_1 = (\alpha, \beta, \gamma)$ and $\mathbf{q}_2 = (z_1, r, z_2)$, then in this case the Jacobian will have the form

$$\vartheta_R = \begin{pmatrix}J_R & \mathbf{0}_{3\times 3}\\ A^{\mathrm{T}}\dfrac{\partial \mathbf{a}}{\partial \mathbf{q}_1} & A^{\mathrm{T}}\dfrac{\partial \mathbf{a}}{\partial \mathbf{q}_2}\end{pmatrix}.$$

Here the matrix J_R is (Chirikjian and Kyatkin, 2001),

$$J_R(A) = \left[\text{vect}\left(A^T\frac{\partial A}{\partial \alpha}\right),\ \text{vect}\left(A^T\frac{\partial A}{\partial \beta}\right),\ \text{vect}\left(A^T\frac{\partial A}{\partial \gamma}\right)\right] = \begin{pmatrix} \sin\beta\sin\gamma & \cos\gamma & 0 \\ \sin\beta\cos\gamma & -\sin\gamma & 0 \\ \cos\beta & 0 & 1 \end{pmatrix}.$$

Due to the block lower diagonal form of this matrix, and the fact that A is a rotation matrix and therefore $\det A = +1$, it is clear that

$$|\det \vartheta_R| = |\det J_R|\left|\det\frac{\partial \mathbf{a}}{\partial \mathbf{q}_2}\right|,$$

and from Eq. A1 it is clear that

$$\frac{\partial \mathbf{a}}{\partial z_1} = \mathbf{e}_3; \quad \frac{\partial \mathbf{a}}{\partial r} = R_3(\alpha)\mathbf{e}_1; \quad \frac{\partial \mathbf{a}}{\partial z_2} = R_3(\alpha)R_1(\beta)\mathbf{e}_3.$$

Therefore, a small computation shows that

$$\left|\det\frac{\partial \mathbf{a}}{\partial \mathbf{q}_2}\right| = \sin\beta,$$

and since $|\det J_R| = \sin\beta$, it follows that

$$|\det \vartheta_R| = \sin^2\beta.$$

Case 2: End-to-line interaction

Performing the multiplications in Eq. 9, we find that

$$H = \begin{pmatrix} R_3(\phi)R_1(\theta)R_3(\alpha)R_1(\gamma) & rR_3(\phi)R_1(\theta)\mathbf{e}_3 + xR_3(\phi)R_1(\theta)R_3(\alpha)\mathbf{e}_1 \\ \mathbf{0}^T & 1 \end{pmatrix}. \tag{A2}$$

The corresponding inverse is

$$H^{-1} = \begin{pmatrix} R_1^T(\gamma)R_3^T(\phi)R_1^T(\theta)R_3^T(\alpha) & -rR_1^T(\gamma)\mathbf{e}_3 \\ \mathbf{0}^T & 1 \end{pmatrix}.$$

Therefore,

$$\left(H^{-1}\frac{\partial H}{\partial \phi}\right)^\vee = \begin{pmatrix} R_1^T(\gamma)R_3^T(\alpha)R_1^T(\theta)\mathbf{e}_3 \\ \cdots\cdots\cdots\cdots \\ rR_1^T(\gamma)R_3^T(\alpha)R_1^T(\theta)E_3R_1(\theta)\mathbf{e}_3 + \\ xR_1^T(\gamma)R_3^T(\alpha)R_1^T(\theta)E_3R_1(\theta)R_3(\alpha)\mathbf{e}_1 \end{pmatrix} = \begin{pmatrix} \sin\theta\sin\alpha \\ \sin\theta\cos\alpha\cos\gamma + \cos\theta\sin\gamma \\ \sin\theta\cos\alpha\sin\gamma - \cos\theta\cos\gamma \\ r\sin\theta\cos\alpha \\ -\sin\theta(r\sin\alpha\cos\gamma + x\cos\alpha\sin\gamma) + x\cos\theta\cos\gamma \\ \sin\theta(r\sin\alpha\sin\gamma - x\cos\alpha\cos\gamma) - x\cos\theta\sin\gamma \end{pmatrix}$$

$$\left(H^{-1}\frac{\partial H}{\partial \theta}\right)^\vee = \begin{pmatrix} R_1^T(\gamma)R_3^T(\alpha)\mathbf{e}_1 \\ \cdots\cdots\cdots\cdots \\ rR_1^T(\gamma)R_3^T(\alpha)E_1\mathbf{e}_3 + \\ xR_1^T(\gamma)R_3^T(\alpha)E_1R_3(\alpha)\mathbf{e}_1 \end{pmatrix} = \begin{pmatrix} \cos\alpha \\ \sin\alpha\cos\gamma \\ \sin\alpha\sin\gamma \\ -r\sin\alpha \\ -r\cos\alpha\cos\gamma + x\sin\alpha\sin\gamma \\ r\cos\alpha\sin\gamma + x\sin\alpha\cos\gamma \end{pmatrix};$$

$$\left(H^{-1}\frac{\partial H}{\partial r}\right)^\vee = \begin{pmatrix} \mathbf{0} \\ \cdots\cdots \\ R_1^T(\gamma)\mathbf{e}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \cdots \\ 0 \\ \sin\gamma \\ \cos\gamma \end{pmatrix};$$

$$\left(H^{-1}\frac{\partial H}{\partial \alpha}\right)^\vee = \begin{pmatrix} R_1^T(\gamma)\mathbf{e}_3 \\ \cdots\cdots\cdots \\ xR_1^T(\gamma)E_3\mathbf{e}_1 \end{pmatrix} = \begin{pmatrix} 0 \\ \sin\gamma \\ \cos\gamma \\ 0 \\ x\cos\gamma \\ -x\sin\gamma \end{pmatrix};$$

$$\left(H^{-1}\frac{\partial H}{\partial x}\right)^\vee = \begin{pmatrix} \mathbf{0} \\ \cdots\cdots \\ \mathbf{e}_1 \end{pmatrix};$$

$$\left(H^{-1}\frac{\partial H}{\partial \gamma}\right)^\vee = \begin{pmatrix} \mathbf{e}_1 \\ \cdots\cdots \\ \mathbf{0} \end{pmatrix}.$$

Stacking these vectors as columns in a 6×6 Jacobian matrix, and taking the determinant, results in

$$|\det \vartheta_R| = r\sin\theta.$$

Case 3: End-to-end interaction

Performing the multiplications in Eq. 11,

$$H = \begin{pmatrix} R_3(\phi)R_1(\theta)R_3(\alpha)R_1(\beta)R_3(\gamma) & rR_3(\phi)R_1(\theta)\mathbf{e}_3 \\ \mathbf{0}^T & 1 \end{pmatrix}. \tag{A3}$$

The corresponding inverse is

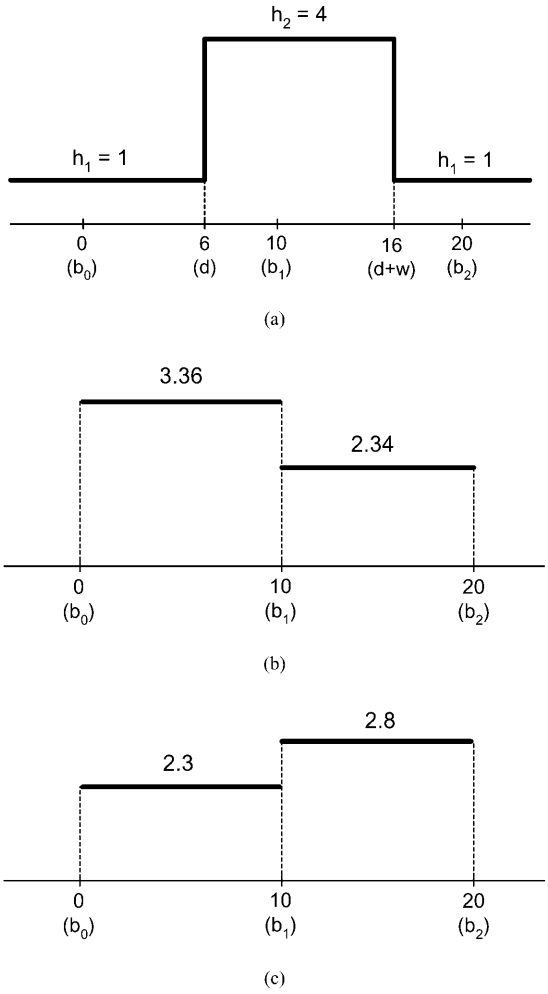

FIGURE A1 Biasing effects of binning before normalizing when $w(\theta) = \sin^2\theta$. (*a*) Ideal histogram. (*b*) Histogram with binning before normalization. (*c*) Histogram with normalization before binning.

$$H^{-1} = \begin{pmatrix} R_3^{\mathrm{T}}(\gamma)R_1^{\mathrm{T}}(\beta)R_3^{\mathrm{T}}(\alpha)R_1^{\mathrm{T}}(\theta)R_3^{\mathrm{T}}(\phi) & -rR_3^{\mathrm{T}}(\gamma)R_1^{\mathrm{T}}(\beta)\mathbf{e}_3 \\ \mathbf{0}^{\mathrm{T}} & 1 \end{pmatrix}.$$

Therefore,

$$\left(H^{-1}\frac{\partial H}{\partial \phi}\right)^{\vee} = \begin{pmatrix} R_3^{\mathrm{T}}(\gamma)R_1^{\mathrm{T}}(\beta)R_3^{\mathrm{T}}(\alpha)R_1^{\mathrm{T}}(\theta)\mathbf{e}_3 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ rR_3^{\mathrm{T}}(\gamma)R_1^{\mathrm{T}}(\beta)R_3^{\mathrm{T}}(\alpha)R_1^{\mathrm{T}}(\theta)E_3R_1(\theta)\mathbf{e}_3 \end{pmatrix}$$

$$= \begin{pmatrix} \sin\theta(\sin\alpha\cos\gamma+\cos\alpha\cos\beta\sin\gamma)+\cos\theta\sin\beta\sin\gamma \\ -\sin\theta(\sin\alpha\sin\gamma-\cos\alpha\cos\beta\cos\gamma)+\cos\theta\sin\beta\cos\gamma \\ \sin\theta\cos\alpha\sin\beta-\cos\theta\cos\beta \\ r\sin\theta(\cos\alpha\cos\gamma-\sin\alpha\cos\beta\sin\gamma) \\ -r\sin\theta(\cos\alpha\sin\gamma+\sin\alpha\cos\beta\cos\gamma) \\ r\sin\alpha\sin\beta\sin\theta \end{pmatrix}$$

$$\left(H^{-1}\frac{\partial H}{\partial \theta}\right)^{\vee} = \begin{pmatrix} R_3^{\mathrm{T}}(\gamma)R_1^{\mathrm{T}}(\beta)R_3^{\mathrm{T}}(\alpha)\mathbf{e}_1 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ rR_3^{\mathrm{T}}(\gamma)R_1^{\mathrm{T}}(\beta)R_3^{\mathrm{T}}(\alpha)E_1\mathbf{e}_3 \end{pmatrix}$$

$$= \begin{pmatrix} \cos\alpha\cos\gamma-\sin\alpha\cos\beta\sin\gamma \\ -\cos\alpha\sin\gamma-\sin\alpha\cos\beta\cos\gamma \\ \sin\alpha\sin\beta \\ -r(\sin\alpha\cos\gamma+\cos\alpha\cos\beta\sin\gamma) \\ r(\sin\alpha\sin\gamma-\cos\alpha\cos\beta\cos\gamma) \\ r\cos\alpha\sin\beta \end{pmatrix};$$

$$\left(H^{-1}\frac{\partial H}{\partial r}\right)^{\vee} = \begin{pmatrix} \mathbf{0} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ R_3^{\mathrm{T}}(\gamma)R_1^{\mathrm{T}}(\beta)\mathbf{e}_3 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0} \\ \cdots\cdots\cdots\cdots \\ \sin\beta\sin\gamma \\ \sin\beta\cos\gamma \\ \cos\beta \end{pmatrix};$$

$$\left(H^{-1}\frac{\partial H}{\partial \alpha}\right)^{\vee} = \begin{pmatrix} R_3^{\mathrm{T}}(\gamma)R_1^{\mathrm{T}}(\beta)\mathbf{e}_3 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \sin\beta\sin\gamma \\ \sin\beta\cos\gamma \\ \cos\beta \\ \cdots\cdots\cdots\cdots \\ \mathbf{0} \end{pmatrix};$$

$$\left(H^{-1}\frac{\partial H}{\partial \beta}\right)^{\vee} = \begin{pmatrix} R_3^{\mathrm{T}}(\gamma)\mathbf{e}_1 \\ \cdots\cdots\cdots\cdots \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \cos\gamma \\ -\sin\gamma \\ 0 \\ \cdots\cdots \\ \mathbf{0} \end{pmatrix};$$

$$\left(H^{-1}\frac{\partial H}{\partial \gamma}\right)^{\vee} = \begin{pmatrix} \mathbf{e}_3 \\ \cdots\cdots \\ \mathbf{0} \end{pmatrix}.$$

Forming the Jacobian matrix and taking its determinant, one finds

$$|\det\vartheta_{\mathrm{R}}| = r^2\sin\beta\sin\theta.$$

A.2 Artifacts resulting from binning before normalizing

As has been explained in the main part of the text, as well as in the articles of Bowie (1997a) and Walther et al. (1998), observed helix-helix packing angle distributions must be normalized by the appropriate geometric factors to obtain true preferences. In the previous Appendix we derived the geometric factors for two packing types not considered in previous works, and verified the case considered by Walther et al. (1998).

In this Appendix the biasing effects of binning before normalizing (which is the computation performed in Bowie, 1997a, and Walther et al., 1998) is examined. A simple closed-form example in one variable illustrates the importance of this effect.

In short, if $f(\theta)$ is an observed angular distribution, and $w(\theta)$ is a geometric weighting factor (e.g., $\sin\theta$ or $\sin^2\theta$), then the function which describes true preferences (in comparison with the uniformly random distribution) is $\rho(\theta)$, where

$$f(\theta)=w(\theta)\rho(\theta). \tag{A4}$$

Therefore, at all points for which $w(\theta)\neq 0$, one can obtain the preference distribution as

$$\rho(\theta)=f(\theta)/w(\theta). \tag{A5}$$

Let us define the histogram of $\rho(\theta)$, which is piecewise constant over equal-sized bins, as

$$[\rho(\theta)]=\sum_{i=0}^{n-1}\rho_i W(\theta,b_i,b_{i+1}), \tag{A6}$$

where

$$\rho_i=\int_{b_i}^{b_{i+1}}\rho(\theta)d\theta \tag{A7}$$

is the constant height of bin i, and $W(\theta, a, b)$ is the window function which is equal to the number 1 on $a\leq\theta\leq b$ and zero otherwise.

Note that Eqs. A5 and A6 were not used in Bowie (1997a) and Walther et al. (1998) to generate histograms. In those works, instead of computing $[\rho(\theta)]=[f(\theta)/w(\theta)]$, the quantity computed was

$$[f(\theta)]/[w(\theta)]\neq[f(\theta)/w(\theta)]. \tag{A8}$$

The central observation of this Appendix is the fact that the above statement is not an equality, and the magnitude of the difference of the two sides in the above ''nonequality'' can be quite dramatic.

Consider the following example, where $w(\theta)=\sin\theta$, and the true underlying angular preference is

$$\rho(\theta)=h_1+(h_2-h_1)W(\theta,d,d+w), \tag{A9}$$

where $h_1=1$, $h_2=4$, $d=6$, and $w=10^\circ$. See Fig. A1 for more information. Also suppose we take the bins to be defined by $b_i=i\times 10^\circ$, and $n=18$ is the number of bins. In this example, the true peak of the preference distribution in Eq. A9 then lies 40% in bin 1, and 60% in bin 2. If one were presented with a plot of this $\rho(\theta)$, one would conclude that its mode is $\sim 12^\circ$.

After binning, the heights of the first two bins are, respectively,

$$\rho_0=\frac{1}{b_1}[h_1 d+h_2(b_1-d)]$$

and

$$\rho_1=\frac{h_2(d+w-b_1)+h_1(b_2-d-w)}{b_2-b_1}.$$

These equations are obtained by simply evaluating Eq. A9 in Eq. A7. This is analogous to what would be computed if each observation of $f(\theta)=\rho(\theta)\sin\theta$ were divided by $\sin\theta$ before being binned. Of course, the preference expressed in $\rho(\theta)$ is not observed directly, and so one must resort to binning, running averages, or kernel-based density estimation methods (Silverman, 1986) to obtain a good estimate of $\rho(\theta)$. Using the binning method and the numerical values in this example, bin 1 of the histogram has a height of 2.3 and bin 2 has a height of 2.8. This proportional splitting of the actual peak which straddles two bins is to be expected. It is simply something that one lives with if one chooses to use histogram methods rather than running averages or kernel methods for density estimation.

However, if one bins first and then normalizes, a very different picture emerges. If we bin first and then normalize, the results for the first two bins are

$$f_0/w_0=\frac{h_1(1-\cos d)+h_2(\cos d-\cos b_1)}{1-\cos b_1}$$

and

$$f_1/w_1=\frac{h_2(\cos b_1-\cos(d+w))+h_1(\cos(d+w)-\cos b_2)}{\cos b_1-\cos b_2}.$$

Using the same numbers as before, the result now is that bin 1 has a value of 2.91 and bin 2 has a value of 2.57. In other words, the results are skewed toward bin 1, even though majority of the true peak should actually be in bin 2.

If we had used $w(\theta)=\sin^2\theta$ instead of $\sin\theta$, the skewing of the histogram toward the first bin would have been even greater: bin 1 has a value of 3.36 and bin 2 has a value of 2.34. This example explains why Walther et al. (1998) observed peaks near the ends of their histograms rather than what we have found. That is, there is no strong preference for helices to interact at $\pm 180^\circ$, but there is a distribution that has a mode in the range 150–170° and tapers to zero as θ approaches $\pm 180^\circ$.

Of course, if a sufficiently large number of sample observations of helix-helix interactions could be observed then bins much finer than 10° could be used, in which case the biasing effects of binning before normalizing would be reduced.

We thank A. Trovato and F. Seno for useful discussions and the anonymous reviewers for their helpful comments.

This work was performed under National Science Foundation grant IIS-0098382. The views expressed are those of the authors alone.

REFERENCES

Abagyan, R. A. 1993. Towards protein-folding by global energy optimization. *FEBS Lett.* 325:17–22.

Adamian, L. J., and J. Liang. 2001. Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J. Mol. Biol.* 311:891–907.

Anfinsen, C. B. 1973. Principles that govern folding of protein chains. *Science.* 181:223–230.

Baldwin, R. L., and G. D. Rose. 1999a. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* 24:26–33.

Baldwin, R. L., and G. D. Rose. 1999b. Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci.* 24: 77–83.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

Bonneau, R., and D. Baker. 2001. Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* 30:173–189.

Bowie, J. U. 1997a. Helix packing angle preferences. *Nat. Struct. Biol.* 4:915–917.

Bowie, J. U. 1997b. Helix packing in membrane proteins. *J. Mol. Biol.* 272:780–789.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.

Cheung, M. S., A. E. Garcia, and J. N. Onuchic. 2002. Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl. Acad. Sci. USA.* 99:685–690.

Chirikjian, G. S., and A. B. Kyatkin. 2001. Engineering Applications of Noncommutative Harmonic Analysis. CRC Press, Boca Raton, FL.

Chothia, C., M. Levitt, and D. Richardson. 1977. Structure of proteins: packing of α-helices and pleated sheets. *Proc. Natl. Acad. Sci. USA.* 74:4130–4134.

Chothia, C., M. Levitt, and D. Richardson. 1981. Helix to helix packing in proteins. *J. Mol. Biol.* 145:215–250.

Creighton, T. E. 1992. Protein Folding. W. H. Freeman, New York.

Crick, F. 1953. The packing of α-helices: simple coiled coils. *Acta Crystallogr.* 6:689–697.

Efimov, A. V. 1979. Packing of α-helices in globular proteins: layer-structure of globin hydrophobic cores. *J. Mol. Biol.* 134:23–40.

Eilers, M., A. B. Patel, W. Liu, and S. O. Smith. 2002. Comparison of helix interactions in membrane and soluble α-bundle proteins. *Biophys. J.* 82:2720–2736.

Finkelstein, A. V., and O. B. Ptitsyn. 1987. Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.* 50:171–190.

Fleishman, S. J., and N. Ben-Tal. 2002. A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane α-helices. *J. Mol. Biol.* 321:363–378.

Fleming, P. J., and F. M. Richards. 2000. Protein packing: dependence on protein size, secondary structure and amino acid composition. *J. Mol. Biol.* 299:487–498.

Lesk, A. M. 2001. Introduction to Protein Architecture. Oxford University Press, Oxford, UK.

Levitt, M., and C. Chothia. 1976. Structural patterns in globular proteins. *Nature.* 261:552–558.

Lin, S. L., C. J. Tsai, and R. Nussinov. 1995. A study of 4-helix bundles—investigating protein-folding via similar architectural motifs in protein cores and in subunit interfaces. *J. Mol. Biol.* 248:151–161.

MacKenzie, K. R., and D. M. Engelman. 1998. Structure-based prediction of the stability of transmembrane helix-helix interactions: the sequence dependence of glycophorin A dimerization. *Proc. Natl. Acad. Sci. USA.* 95:3583–3590.

Maiorov, V. N., and G. M. Crippen. 1992. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* 227:876–888.

Miyazawa, S., and R. L. Jernigan. 1985. Estimation of effective inter-residue contact energies from protein crystal-structures-quasi-chemical approximation. *Macromolecules.* 18:534–552.

Miyazawa, S., and R. L. Jernigan. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256:623–644.

Murzin, A. G., and A. V. Finkelstein. 1988. General architecture of the α-helical globule. *J. Mol. Biol.* 204:749–769.

Reddy, B., and T. Blundell. 1993. Packing of secondary structural elements in proteins: analysis and prediction of interhelix distance. *J. Mol. Biol.* 233:464–479.

Richmond, T. J., and F. M. Richards. 1978. Packing of α-helices: geometrical constraints and contact areas. *J. Mol. Biol.* 119:537–555.

Robinson, C. R., and S. G. Sligar. 1993. Electrostatic stabilization in four-helix bundle proteins. *Protein Sci.* 2:826–837.

Silverman, B. W. 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, UK.

Sippl, M. J. 1995. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* 5:229–235.

Skolnick, J., and A. Kolinski. 1999. Monte Carlo approaches to the protein folding problem. *Adv. Chem. Phys.* 105:203–242.

Srinivasan, R., and G. D. Rose. 1995. LINUS—a hierarchical procedure to predict the fold of a protein. *Prot. Struct. Funct. Genet.* 22:81–99.

Trovato, A., and F. Seno. 2003. A new perspective on the analysis of helix-helix packing preferences in globular proteins. arXiv:cond-mat/0304429 v1 18 April2003.

Walther, D., F. Eisenhaber, and P. Argos. 1996. Principles of helix-helix packing in proteins: the helical lattice superposition model. *J. Mol. Biol.* 255:536–553.

Walther, D., C. Springer, and F. E. Cohen. 1998. Helix-helix packing angle preferences for finite helix axes. *Prot. Struct. Funct. Genet.* 33:457–459.

Wang, G., and R. L. Dunbrack. 2003. Culling the PDB by resolution and sequence identity. http://www.fccc.edu/research/labs/dunbrack/pisces/culledpdb.html.

Weaver, D. L. 1992. Hydrophobic interaction between globin helices. *Biopolymers.* 32:477–490.

Weiner, P. K., and P. A. Kollman. 1981. AMBER: assisted model building with energy refinement. A general program for modeling molecules and their interactions. *J. Comput. Chem.* 2:287–303.

Zhou, F. X., M. J. Cocco, W. P. Russ, A. T. Brunger, and D. M. Engelman. 2000. Interhelical hydrogen bonding drives strong interactions in membrane proteins. *Nat. Struct. Biol.* 7:154–160.